

Elements of Coevolution in Biological Sequences

Olivier Rivoire

CNRS, LIPhy UMR 5588, Grenoble, France

Studies of coevolution of amino acids within and between proteins have revealed two types of coevolving units: coevolving contacts, which are pairs of amino acids distant along the sequence but in contact in the three-dimensional structure, and sectors, which are larger groups of structurally connected amino acids that underly the biochemical properties of proteins. By reconciling two approaches for analyzing correlations in multiple sequence alignments, we uncover a new class of coevolving units, called 'sectons'. Sectons provide a conceptual link between coevolving contacts and sectors. The methods and results that we present are general, and relevant beyond protein structures. This generality is illustrated with an analysis of the co-occurrence of orthologous genes in bacterial genomes.

PACS numbers: 02.50.Sk, 87.14.E-, 87.15.Qt, 87.18.Wd

The structural and functional properties of proteins emerge from interactions between their amino acids. During evolution, these interactions constrain the substitutions of amino acids that may happen. Sequences resulting from multiple independent evolutionary trajectories reflect these constraints, and therefore contain information about the organization of interactions within proteins. Such sequences are now made available by DNA sequencing technology, which provides thousands of protein sequences that have diverged independently and under similar selective pressures from a common ancestral sequence.

These protein sequences are commonly collected into multi-sequence alignments on the basis of their sequence similarity. Alignments for over 10^4 families of protein domains are for instance available in the Pfam database [1]. Formally, an alignment is described by a $M \times L \times A$ binary array x_{si}^a , where $x_{si}^a = 1$ indicates that sequence $s \in \{1, \dots, M\}$ has amino acid $a \in \{1, \dots, A\}$ at position $i \in \{1, \dots, L\}$, with $x_{si}^a = 0$ otherwise; some positions contain gaps, inserted to ensure an optimal alignment of sequences, which are represented by $x_{si}^a = 0$ for every amino acid $a = 1, \dots, A$, where $A = 20$. Typical numbers are $M \sim 10^2$ - 10^4 for the number of sequences, $L \sim 10^2$ - 10^3 for the length of the alignment.

The pattern of amino acid interactions may be inferred from the statistical correlations between pairs of positions in the alignment. Analyses of these correlations are complicated by several factors: (i) proteins are gathered in an alignment based on sequence similarity, with no guarantee to have been subject to common selective constraints; (ii) sequences are not sampled independently during evolution, but through a branching process, which introduces a sampling bias; (iii) the information content of the alignment, $\sim ML \log_2 A \sim 10^5$ - 10^7 bits, is small compared to the number $\sim A^2 L^2 / 2 \sim 10^6$ - 10^8 of continuous parameters defining the correlations between every pair of amino acids, which implies a severe under-sampling; (iv) two positions may be correlated while not directly interacting, reflecting a fundamental difference

between interactions and correlations.

Standard statistical analyses assimilate the observed samples to an asymptotically large number of independently and identically distributed random variables. Points (i), (ii) and (iii) violate each of these assumptions, while point (iv) suggests that, even in absence of bias, further processing is required to infer interactions from correlations.

Many approaches have been proposed to tackle these challenges [2]. Recently, two methods have been developed, each rooted in a different concept of statistical mechanics, and each providing results of different nature. In an extension of an approach called Statistical Coupling Analysis (SCA) [3], an application of concepts from random matrix theory [4] to address (iii) has revealed collective modes of coevolution named 'sectors' [5]. A protein sector consists of ~ 15 -30 positions that are connected in the three-dimensional structure, and experiments indicate that each sector controls independently a biochemical property of the protein [5]. In a different approach called Direct Coupling Analysis (DCA) [6], the problem (iv) of inferring interactions from correlations was formulated and solved as a problem of inverse statistical mechanics, leading to the inference of a large number of pairs of positions in contact in the three-dimensional structure [7].

The two approaches, SCA and DCA, differ in their principles as well as in their results. In a comparison using a common alignment, the contacts inferred by DCA seemed to bear no relation to the sectors identified by SCA [7]. We provide here a rationale for these apparently dissonant results. We show (1) how the two approaches expose two aspects of a common pattern of amino acid interactions; (2) how new units of coevolution can be defined, which underly the contacts inferred by DCA. We name these elementary units 'sectons', and illustrate their relation to sectors with the trypsin family of enzymes.

Our arguments are general, and the notion of sectons relevant not only to other protein families, but also to

datasets of different nature. We demonstrate it by applying the same methods to the co-occurrence of orthologous genes in bacterial sequences, also known as their phylogenetic profile [8]. We show (3) how sectors can be identified at the scale of the genome, which define elementary units of co-functional genes.

We begin with an analysis of coevolution in the trypsin family of protein sequences, using the alignment from Pfam [1], which contains nearly 15000 sequences. SCA and DCA both start from the same correlation matrix C_{ij}^{ab} , which reports the coevolution of amino acid a at position i with amino acid b at position j . Prior to defining this matrix, some steps must be taken to clean the alignment from positions with excessive gaps and mitigate the effects of (i) and (ii) by weighting differentially the contributions of the various sequences. These steps are straightforward but essential, and may be common for both approaches (all details are provided as Supplementary Material [9]). The outcome is the definition of f_i^a , the frequency of amino acid a at position i , and f_{ij}^{ab} , the joint frequency of (a, b) at the pair of positions (i, j) . These frequencies define the correlation matrix C_{ij}^{ab} as

$$C_{ij}^{ab} = f_{ij}^{ab} - f_i^a f_j^b. \quad (1)$$

SCA aims at identifying groups of positions under selection for a common functional property, based on two principles: the conservation of amino acids involved in the function, and their correlations induced by cooperative interactions. SCA takes a heuristic approach to combine these two principles by weighting the correlations C_{ij}^{ab} with a measure of amino acid conservation W_i^a , defined by

$$W_i^a = \left| \ln \left(\frac{f_i^a(1 - q^a)}{(1 - f_i^a)q^a} \right) \right|, \quad (2)$$

where $q^a = \sum_{i=1}^L f_i^a / L$ is the mean frequency of amino acid a ; thus, the more f_i^a deviates from q^a and approaches 1, the larger W_i^a is. Taking $W_i^a W_j^b C_{ij}^{ab}$ defines a conservation-weighted correlation matrix, which is reduced to a $L \times L$ correlation matrix between positions as follows [3, 10]:

$$C_{ij} = \sqrt{\sum_{a,b} (W_i^a W_j^b C_{ij}^{ab})^2}. \quad (3)$$

Following an approach first proposed to infer business sectors from correlations between financial time series [4], protein sectors are identified from the top eigenvectors of this matrix [5]. This definition is facilitated by rotating the top eigenvectors into maximally independent components, using independent component analysis (ICA) [11, 12]. Sectors can be defined from the top k_{top} components $V^{(k)}$ as $\mathcal{S}_k = \{i : V_i^{(k)} > \epsilon\}$. Here, we take $k_{\text{top}} = 4$ and $\epsilon = 0.1$, but the results are insensitive to

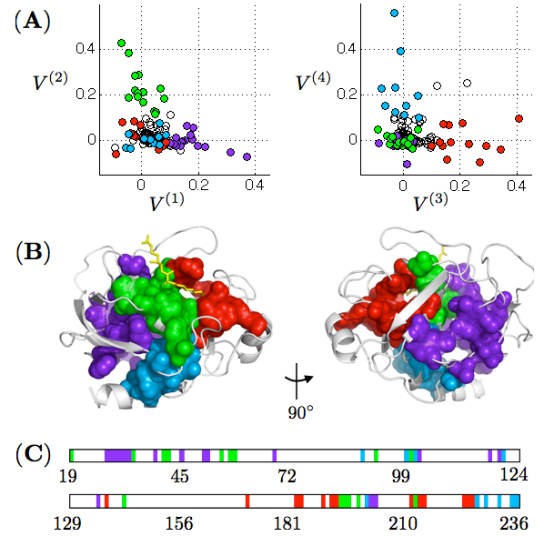


FIG. 1: Protein sectors in the trypsin family, as inferred from the Pfam alignment PF00089 [1] – (A) Projections of the positions i along the vectors $V^{(k)}$ obtained by rotating by ICA the top 4 eigenvectors of C_{ij} . Sector k is defined by the positions i with $V_i^{(k)} > \epsilon$ and $V_i^{(\ell)} < \epsilon$ for $\ell \neq k$, with $\epsilon = 0.1$. (B) Location of the sectors on a three-dimensional structure of trypsin [13]. (C) Location of the sectors along the sequence (cut in two for readability), with non-sector positions in white (numbering system of bovine chymotrypsin).

these exact values; increasing k_{top} or decreasing ϵ in fact provides complementary information [9]. The sectors are represented with different colors in Fig. 1 (for simplicity, these sectors do not include the positions i with $V_i^{(k)} > \epsilon$, for multiple k [9]).

As shown in Fig. 1, each sector forms a connected group of positions on the three-dimensional structure, despite not necessarily consisting of consecutive positions along the sequence. Sectors have no sharp boundaries, but are typically organized into an onion-like hierarchy, with the core of sector k consisting of positions i with largest $V_i^{(k)}$, and layers associated with decreasing values of $V_i^{(k)}$ [9]. Three sectors were previously inferred for the same protein family using an alignment about 10 times smaller [5, 9]: two of these sectors, respectively associated with enzymatic activity and specificity, correspond to the green and red sectors in Fig. 1; the third one, which had the peculiarity of a disconnected core, and which correlated experimentally with stability, is now partly spread over two new sectors, whose functional role remains to be characterized.

In contrast to SCA, DCA aims at identifying structural contacts between positions by inferring direct interactions from indirect correlations. It proceeds by a mapping to the problem of reconstructing the couplings J_{ij}^{ab} of a $(A + 1)$ -state Potts model given its correla-

tions C_{ij}^{ab} . Solving exactly this problem is computationally prohibitive, but mean-field approximations provide a range of alternatives [14]. The simplest of these approximations consists in taking $J = -C^{-1}$, where C , given by Eq. (1), is treated as a $(AL) \times (AL)$ matrix. As C has typically rank $< M$, it is not invertible but it can be regularized to

$$\bar{C} = \lambda C + (1 - \lambda)Q, \quad Q_{ij}^{ab} = q^a(\delta^{ab} - q^b)\delta_{ij}, \quad (4)$$

where Q represents a background expectation, so that we can take $J = -\bar{C}^{-1}$ (with $\lambda = 1/2$ here as in Ref. [7]). Regularization is not necessary for SCA, but substituting \bar{C}_{ij}^{ab} for C_{ij}^{ab} in Eq. (3) does not alter the definition of sectors [9].

The couplings J_{ij}^{ab} define for $i \neq j$ a model for the distribution of amino acids at every pair of positions ij ,

$$g_{ij}^{ab} = \exp(J_{ij}^{ab} + h_i^a + h_j^b + h_0), \quad (5)$$

where h_i^a , h_j^b , h_0 are uniquely determined by requiring that $\sum_b g_{ij}^{ab} = \bar{f}_i^a$ with $\bar{f}_i^a = \lambda f_i^a + (1 - \lambda)q^a$, and by imposing an overall normalization. From g_{ij}^{ab} , a matrix of 'direct information' [6] is defined by

$$\mathcal{D}_{ij} = \sum_{a,b} g_{ij}^{ab} \ln [g_{ij}^{ab} / (\bar{f}_i^a \bar{f}_j^b)]. \quad (6)$$

As shown in Ref. [7], many of the pairs ij with top values \mathcal{D}_{ij} are in contact in the three-dimensional structure, to the extent that these contacts provide sufficient constraints to infer the structure [15] (contacts are here defined by a distance < 8 Å).

In this work, we follow Ref. [7] in defining \mathcal{D}_{ij} by Eq. (6), but we truncate it before analyzing it by ICA as we did for \mathcal{C}_{ij} . Many of the top pairs in terms of \mathcal{D}_{ij} are indeed induced by the presence of stretches of gaps and are therefore consecutive along the sequence; to focus on non-trivial contacts, we substitute \mathcal{D}_{ij} with $\tilde{\mathcal{D}}_{ij}$, where $\tilde{\mathcal{D}}_{ij} = \mathcal{D}_{ij}$ if $|i - j| > \Delta$, and 0 otherwise. We take here $\Delta = 5$, but other values give consistent results.

As \mathcal{C}_{ij} , the matrix $\tilde{\mathcal{D}}_{ij}$ can be analyzed by extracting its top eigenvectors and rotating them by ICA. Remarkably, this leads to a large number (~ 100) of independent components, each localized on a small group of 2 to 5 positions. Fig. 2 shows the first 24 such groups using $k_{\max} = 120$ and $\epsilon = 0.2$, but similar results are obtained for a range of values of k_{\max} and ϵ [9]. Out of the 24 sections shown in Fig. 2, only one, the 23rd, contains a position that is not in contact with the others. We call these small structural elements 'protein sections'. Sections underlie the contacts inferred by DCA, but, in a majority of cases, they consist of more than two positions. Only few sections are well-recognized structural or functional units: for the trypsin family, the top 6 sections thus include 4 disulfide bonds (pairs of cysteines forming covalent bonds) [20], and the 'catalytic triad', a

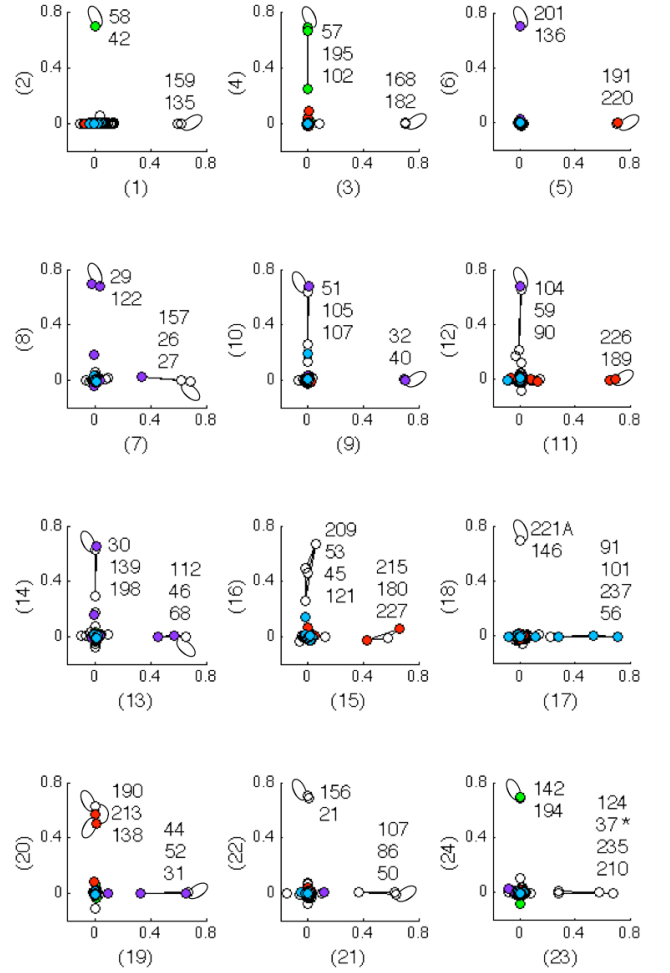


FIG. 2: Top protein sections in the trypsin family – Each graph is a projection of the positions along $(U^{(k)}, U^{(k+1)})$, the components of order k and $k + 1$ obtained by rotating by ICA the top eigenvectors of the truncated matrix of direct information $\tilde{\mathcal{D}}_{ij}$. Sections are defined by $s_k = \{i : U_i^{(k)} > \epsilon\}$, with $\epsilon = 0.2$. The labeling of positions follows the numbering system of bovine chymotrypsin (note that in several instances positions appear as superimposed). Positions in a section are joined by a line when their distance is < 8 Å; by this criterion, all sections shown here are structurally connected, except for s_{23} , where position 37 is distant from the others. The colors are from Fig. 1, showing that sections follow the decomposition into sectors (non-sector positions are in white). Sections s_2, s_3, s_5, s_6 are disulfide bonds, and s_4 is the catalytic triad.

combination of three amino acids shared by several other protein families [16]. The other sections have not been described previously and characterizing their structural and/or functional roles is a clear experimental challenge.

Formally, sectors and sections can be associated with two exclusive parts of the spectrum of a common correlation matrix, \bar{C} [9]. They are however not unrelated: sections are found within or outside sectors, but not be-

tween sectors, as seen in Fig. 2. This absence of inter-sector sections reflects the evolutionary independence of sectors. Sections thus define elementary units of coevolution that are consistent with the overall decomposition into sectors.

Sections are similarly found in other protein families [10], but the concept is not limited to protein structures. As another example involving biological sequences, but at another scale and with data of different nature, we consider here the problem of inferring the functional couplings between genes in a genome. A first-order approach to this problem is to study the co-occurrence of genes in a large number of genomes, with, as raw data, an $M \times L$ binary matrix x_{si} , where $x_{si} = 1$ indicates that gene i is present in the genome of species s , and 0 that it is absent ($A = 1$ in this case). Building such a dataset requires mapping corresponding genes across genomes: we rely here on the partition of bacterial genes into clusters of orthologous genes (COGs) [17], to obtain a dataset consisting of $M \simeq 10^3$ genomes and $L \simeq 1.5 \cdot 10^3$ orthologous classes of genes [9].

The same methods lead to the identification of many genomic sections, the first of which are displayed in Fig. 3. Many of these sections have a clear structural or functional interpretation: several are different subunits of a same protein complex, and others are known to be involved in a common function [9]. Some sections, however, cannot be easily interpreted, as they involve currently uncharacterized proteins: for such cases, our results predict new functional relations. Genomic sections, involving larger groups of genes, may be defined as well, although their significance is more difficult to assess [9].

The methods allow for many variations, and we presented them only in their simplest instantiations. Many alternative measures for scoring coevolution are for instance possible. Among them, a correlation matrix as in Eq. (3) but without weights ($W_i^a = 1$ for all i, a), or the 'mutual information' [18], can be analyzed along the same lines: interestingly, these matrices yield coevolving elements that are intermediate between sectors and sections [9]. More sophisticated methods should allow for better characterization of the patterns of coevolution in biomolecules, in particular to account for their hierarchical organization. New approaches will for instance integrate phylogenetic data, or take advantage of recent progress in inverse statistical mechanics [14] and high-dimensional statistics [19].

In conclusion, we extended the previously reported decomposition of proteins into protein sectors [5] to a decomposition into a hierarchy of structural elements, with a new class of coevolving units, sections, at its basis. Sections, rather than pairs of contacting positions, may be the relevant target for methods aiming at inferring protein structures from multiple sequence alignments. We also provided evidence that the same pattern of coevolution extends beyond protein structures, to the

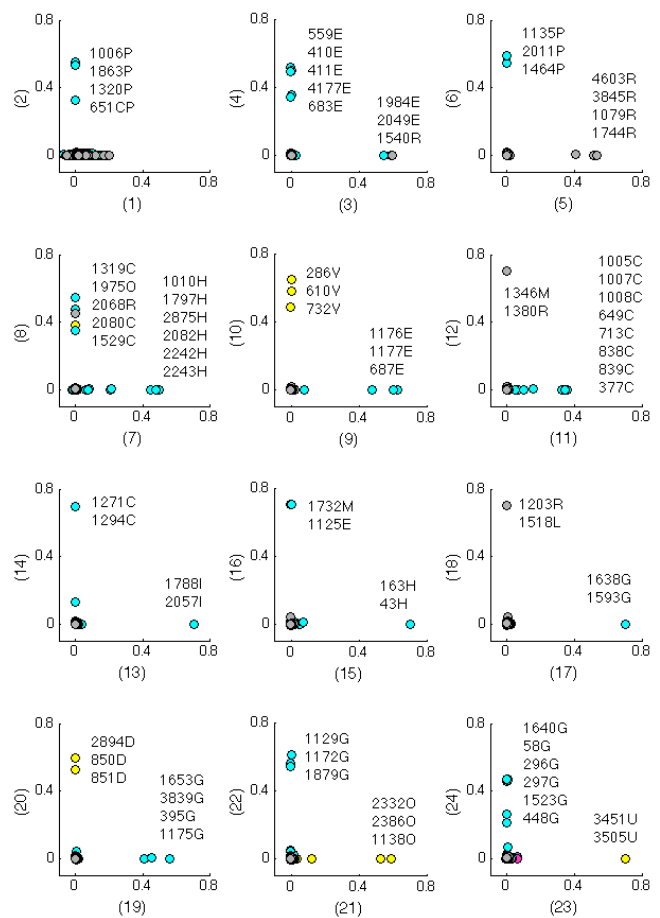


FIG. 3: Top genomic sections in bacteria – As Fig. 2, but for the co-occurrence of orthologous genes in bacterial genomes. Each dot represents a cluster of orthologous genes (COG), with a colors associated with its functional class: cyan for metabolism, yellow for cellular processes, magenta for information processing, and gray for poorly characterized genes [17]. The sections typically comprise genes from a common functional class, and often even subclass, which is indicated by the last letter labeling the COGs.

scale of the genome. Only a small fraction of protein or genomic sections have been previously recognized as fundamental units by means of other approaches. Characterizing generally the structural, functional and evolutionary roles of sections is an open problem that extends beyond the scope of statistical studies of sequence data. In particular, experiments are needed to assess the extent to which statistical patterns of coevolution, inferred from a collection of sequences, are reflected in individual biomolecules.

Acknowledgements – I thank L. Colwell, B. Houchmandzadeh, I. Junier, S. Leibler, R. Ranganathan, K. Reynolds, T. Tesileanu for discussions and comments. This work is supported by ANR grant 'CoevolInterProt'.

-
- [1] M. Punta, P. C. Coghill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, et al., *Nucl Acids Res* **40**, D290 (2011).
 - [2] D. S. Horner, W. Pirovano, and G. Pesole, *Briefings Bioinfo* **9**, 46 (2007).
 - [3] S. Lockless and R. Ranganathan, *Science* **286**, 295 (1999).
 - [4] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, T. Guhr, and H. E. Stanley, *Phys. Rev. E* **65**, 066126 (2002).
 - [5] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, *Cell* **138**, 774 (2009).
 - [6] M. Weigt, R. White, H. Szurmant, J. Hoch, and T. Hwa, *Proc Nat Acad Sci* **106**, 67 (2009).
 - [7] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, *Proc. Natl. Acad. Sci.* **108**, E1293 (2011).
 - [8] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates, *Proc. Natl. Acad. Sci.* **96**, 4285 (1999).
 - [9] See Supplementary Material.
 - [10] O. Rivoire and R. Ranganathan, (in preparation).
 - [11] A. Bell and T. Sejnowski, *Neural Comp.* **7**, 1129 (1995).
 - [12] R. G. Smock, O. Rivoire, W. P. Russ, J. F. Swain, S. Leibler, R. Ranganathan, and L. M. Gierasch, *Mol. Syst. Bio.* **6**, 1 (2010).
 - [13] A. Pasternak, D. Ringe, and L. Hedstrom, *Prot. Science* **8**, 253 (1999).
 - [14] Y. Roudi, E. Aurell, and J. A. Hertz, *Frontiers Comput. Neuro.* **3** (2009).
 - [15] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, *PLoS ONE* **6**, e28766 (2011).
 - [16] J. M. Berg, J. L. Tymoczko, and L. Stryer, *Biochemistry*, chap. 9 (W. H. Freeman & Cie, 2007), 6th ed.
 - [17] R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin, *Nucl. Acids Res.* **28**, 33 (2000).
 - [18] B. Korber, R. Farber, D. Wolpert, and A. Lapedes, *Proc. Natl. Acad. Sci.* **90**, 7176 (1993).
 - [19] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data* (Springer, 2011).
 - [20] Rat trypsin has a total of 6 disulfide bonds but the other 2 are not conserved in the family, with frequencies < 5%.

SUPPLEMENTAL MATERIAL

Preprocessing of the alignment

As input for the identification of sectors and sections in the trypsin family, we downloaded the full alignment PF00089 from Pfam (version 26.0, Nov. 2011) [1]. This alignment contains $M_0 = 14720$ sequences. It is represented by an array x_{si}^a where s labels the sequences (row in the alignment), i the positions (columns) and a is a number between 1 and 20 (each number is associated with one of the 20 amino acids); $x_{si}^a = 1$ indicates that sequence i has amino acid a at position i , and $x_{si}^a = 0$ otherwise. As a reference for truncating the alignment and comparing to structural data, we used the sequence and structure of rat trypsin, chain E of the PDB id 3TGI in the Protein Data Base [13], which consists of $L_0 = 223$ positions.

To clean the alignment from an excess of gaps, the following operations were performed:

- (1) Truncation of positions based on the reference sequence. As the alignment does not contain the last 7 positions of the reference sequence, this step leaves $L_1 = 216$ positions.
- (2) Removal of sequences with a fraction of gaps exceeding $\gamma_{\text{seq}} = 0.2$, or with a sequence similarity to the reference sequence below $s_{\text{min}} = 0.2$, where sequence similarity is defined by

$$S_{rs}^{(1)} = \frac{1}{L_1} \sum_{i,a} x_{ri}^a x_{si}^a. \quad (7)$$

This step leaves $M = 9589$ sequences.

- (3) Removal of positions with a fraction of gaps exceeding $\gamma_{\text{pos}} = 0.2$. The frequencies of gaps are computed with sequence weights as defined by Eq. (8), using $S_{rs}^{(1)}$. This step leaves $L = 204$ positions.

The parameters $\gamma_{\text{seq}} = 0.2$, $\gamma_{\text{pos}} = 0.2$ and $s_{\text{min}} = 0.2$ are chosen to mitigate the effects of gaps, but the results are not sensitive to their exact values. A more in-depth analysis of the structure of sequence correlations can reveal further information, and may suggest the removal of additional sequences, but this analysis is beyond the scope of the present study [10].

Sequence-weighted frequencies

Following Ref. [6], the uneven sampling of sequences is alleviated by introducing sequence weights defined by

$$w_s \equiv \frac{\nu_s^{-1}}{\sum_r \nu_r^{-1}}, \quad \text{with} \quad \nu_s \equiv |\{r : S_{rs} > \delta\}|, \quad (8)$$

i.e., ν_r counts the number of sequences r within distance δ of sequence s , where the distance between two sequences is the fraction of amino acids by which they differ:

$$S_{rs} = \frac{1}{L} \sum_{i,a} x_{ri}^a x_{si}^a. \quad (9)$$

$M' = \sum_r \nu_r^{-1}$ can be interpreted as an effective number of sequences in the alignment. Here we take $\delta = 0.8$, which results in $M' \simeq 4600$ effective sequences for the trypsin alignment.

The sequence weights are used to define frequencies as

$$f_i^a \equiv \sum_s w_s x_{si}^a, \quad f_{ij}^{ab} \equiv \sum_s w_s x_{si}^a x_{sj}^b. \quad (10)$$

Rotation by independent component analysis (ICA)

Different implementations of ICA use different measures of independence and different algorithms for optimizing them. Here, we use one of the simplest implementations of ICA, proposed by Bell and Sejnowski [11], with modifications introduced by Amari [S. Amari, A. Cichocki, and H. H. Yang. A new learning algorithm for blind signal

separation. In D. Touretzky, M. Mozer, and M. Hasselmo, editors, *Advances in neural information processing systems*, volume 8, pages 757-763, Cambridge MA, 1996. MIT Press.]. We take as input the top k eigenvectors of the correlation matrix \mathcal{C}_{ij} or $\tilde{\mathcal{D}}_{ij}$, which we concatenate in a $k_{\text{top}} \times L$ matrix Z (at variance with usual implementations of ICA using as input the dataset X). The algorithm iteratively updates an unmixing matrix W , starting from the $k_{\text{top}} \times k_{\text{top}}$ identity matrix $W_0 = I_{k_{\text{top}}}$, with increments ΔW given by

$$\Delta W = \eta \left(I_{k_{\text{top}}} + \left(1 - \frac{2}{1 + \exp(-WZ)} \right) (WZ)^\top \right) W. \quad (11)$$

The parameter η is a learning rate that has to be sufficiently small for the iterations to converge; in this study, $\eta = 10^{-4}$ led to convergence after 10^4 iterations in every case.

The independent components $V^{(k)}$ (or $U^{(k)}$) are obtained by applying W to the eigenvectors in Z . To set their overall scale and sign, we normalize them to unit length ($\sum_i (V_i^{(k)})^2 = 1$) and orient them so that the position i with largest $|V_i^{(k)}|$ satisfies $V_i^{(k)} > 0$. The order of the independent components, which is generally not prescribed by ICA, is here well defined by the algorithm and is related to the order of the principal components.

Threshold k_{top} in defining sectors

The spectrum of \mathcal{C}_{ij} , displayed in Fig. S1, indicates that between 4 and 8 eigenvalues are emerging from a bulk of small eigenvalues. This estimation is confirmed by comparing with the spectra of randomized alignments, where the amino acids are drawn independently at each position i according the frequencies f_i^a so as to remove the correlations but preserve the distribution of amino acids at each position.

In the main text, we presented the results when selecting $k_{\text{top}} = 4$ modes. A smaller number of components may prevent the discrimination between sectors, as shown in Fig. S2 where taking $k_{\text{top}} = 3$ causes the red and blue sectors to appear along a same component. Reciprocally, a larger number of components may lead to the splitting of a sector into disconnected subsets, as shown in Fig. S3 where $k_{\text{top}} = 5$ decomposes the purple sector along two components. In this case, the two components do not define two new sectors, but rather indicate a partition of the sector into a core and a periphery, as shown in Fig. S4. As increasing values of k_{top} are considered, sectors break up successively into smaller components, as seen in Fig. S5 with $k_{\text{top}} = 24$.

Threshold ϵ in defining sectors

Besides the threshold k_{top} for the eigenvalues, the definition of sectors involves a threshold ϵ determining which positions contribute significantly to each component. Here again, ϵ can be estimated from a comparison with randomized alignments, but it is more interesting to notice that several values of ϵ are consistent with structurally connected sectors. Varying ϵ thus defines a hierarchy of structurally connected positions, from the core of the most conserved positions for large ϵ to a periphery of less conserved positions for smaller ϵ .

As an illustration of this feature, we show in Fig. S6 how the connectivity of each sector, measured by the relative size of its largest structurally connected subset, varies with ϵ . With the possible exception of the cyan sector, the sectors are found to be significantly structurally connected for nearly all values of ϵ . The significance of this finding is assessed by comparing with randomly-formed groups of positions, or with the positions ordered by their overall degree of conservation.

Intersections between sectors

In Fig. 1, we eluded the discussion of positions at the intersection between different $\mathcal{S}_k = \{i : V_i^{(k)} > \epsilon\}$ by excluding them from the definition of sectors. These few positions are however structurally meaningful as well: Fig. S7 shows that they are always found at the periphery of one of the two sectors, and, in several instances, at the structural interface between them.

Composition of sectors

Fig. S8 reports the exact composition of the 4 sectors defined in the main text. The green and red sectors have very significant overlap with the green and red sectors previously defined in Ref. [5] using a smaller alignment. On the other hand, the purple and cyan sectors have only limited overlap with the blue sector defined in this previous study. As a visual representation of the relation between these two definitions, Fig. S9 reproduces Fig. 1, but with colors corresponding to the 3 sectors defined in Ref. [5].

Contacts and sections

In computing the matrix of direct information \mathcal{D}_{ij} we follow the DCA of Ref [7], with only minor differences: (1) we trim the alignment from sequences with significant dissimilarity to the reference sequence, as part of the preprocessing steps; (2) we use background frequencies q^a computed from averages over the alignment, rather than $q^a = 1/(A + 1)$, for consistency with SCA, and with no major incidence; (3) we regularize/shrink the covariance matrix C instead of using pseudo-counts for the frequencies, which has also minor consequences. Finally, we truncate \mathcal{D}_{ij} from its diagonal, which has no incidence on the prediction of contacts when ranking pairs by their value of direct information if restricting to non-trivial contacts, defined as those at distance $> \Delta = 5$ along the chain. Fig. S10 reports the performance of DCA for predicting contacts for the alignment under study: with only one exception, all top 33 predicted and non-trivial contacts are actual contacts. Fig. S11 gives the list of the top 80 non-trivial predictions of contacts and indicates the section to which they belong. Fig. S12 shows the top 24 sections of Fig. 2 on the three-dimensional structure of trypsin.

Fig. S14 shows that $\sim 90\%$ of the top 60 sections are structurally connected. It also shows that, as for sectors, the results are insensitive to the choice of the threshold ϵ used in defining the positions contributing significantly to a component. In all figures involving sections, we use $k_{\text{top}} = 120$, to be able to consider a large number of sections, but the results are here also insensitive to this exact value. Finally, Fig. S13 gives the composition of the top 120 sections and indicates which are connected.

Not truncating \mathcal{D}_{ij} leads to the sections shown in Fig. S15. Many of these sections consist of consecutive positions. These trivial sections are induced by gaps, which tend to be consecutive along the sequence (this feature is partly a consequence of the multiple sequence alignment algorithm, which have a penalty for opening new gaps). Truncating \mathcal{D} (or equivalently J) is a simple if not optimal way of getting rid of these trivial correlations.

Orthogonal decomposition of the correlations

To show how sectors and sections can be derived from two distinct parts of a common correlation matrix, we take here the regularized correlation matrix \bar{C} and decompose it into two orthogonal parts \bar{C}^+ and \bar{C}^- . More precisely, if $\bar{C} = \sum_k |k\rangle \lambda_k \langle k|$ denotes the spectral decomposition of \bar{C} in the bra-ket notation, with $\lambda_1 \geq \dots \lambda_L$, we define

$$\bar{C}^+ = \sum_{k \leq k^*} |k\rangle \lambda_k \langle k|, \quad \text{and} \quad \bar{C}^- = \sum_{k > k^*} |k\rangle \lambda_k \langle k|. \quad (12)$$

We then apply SCA to \bar{C}^+ instead of C , and DCA to $J^- = -\sum_{k > k^*} |k\rangle \lambda_k^{-1} \langle k|$ instead of $J = -\bar{C}^{-1}$. For $k^* = 20$, Fig. S16 shows that the same sectors are recovered from C and \bar{C}^+ , and Fig. S17 that the same sections are recovered from C and \bar{C}^- .

Random matrix theory indicates that both ends of the spectra of under-sampled empirical covariance matrices are statistically significant. The few top modes support the sectors. The bottom modes are essential for defining sections (and therefore for inferring contacts by DCA) but they are not sufficient: sections are not exclusively associated with the bottom modes of \bar{C}_{ij}^{ab} , as also suggested by Fig. S5 and S18 (but they are, by definition, associated with the top modes of $\tilde{\mathcal{D}}_{ij}$). Thus, the number of modes included in the definition of \bar{C}^- cannot be reduced significantly without altering the nature of the sections recovered from it, even though most of the $20L = 4080$ modes of \bar{C} may not be statistically significant.

Unweighted correlations and mutual information

The operations applied to \mathcal{C}_{ij} and $\tilde{\mathcal{D}}_{ij}$ to respectively obtain sectors and sectons, i.e., extraction and rotation by ICA of the top eigenvectors, can be applied to other measures of correlations. In particular, it can be applied to the matrix of unweighted correlations \mathcal{C}_{ij}^0 obtained from Eq. (3) by taking flat positional weights $W_i^a = 1$ for all i, a . Groups of coevolving positions of size intermediate between sectors and sectons, and consistent with both, are thus defined, as shown in Fig. S18.

The matrix of mutual information \mathcal{M}_{ij} can be analyzed similarly. This matrix is defined by

$$\mathcal{M}_{ij} = \sum_{a=0}^{20} f_{ij}^{ab} \ln \frac{f_{ij}^{ab}}{f_i^a f_j^b}, \quad (13)$$

with $a = 0$ corresponds to a gap, so that $f_i^0 = 1 - \sum_{a=1}^{20} f_i^a$ and similarly for f_{ij}^{0b} , f_{ij}^{a0} , and f_{ij}^{00} . Fig. S19(A) shows that its top components do report structurally connected positions, but most of them are trivial, i.e., consist of consecutive positions along the sequence. As for direct information, we can however truncate a band along the diagonal of this matrix to obtain non-trivial groups of correlated positions. These groups of coevolving positions are again both structurally connected and consistent with the decomposition into sectors and sectons, as shown in Fig. S19(B).

Co-occurrence of orthologous genes in bacterial genomes

Sequenced bacterial genomes and COG annotations were downloaded from NCBI. The initial dataset contained $M_0 = 1432$ genomes and $L_0 = 4467$ COGs.

The following cleaning steps were conducted:

- (1) Removal of 'exceptional' genomes with size below 500 kpb or with no less than 60 % of genes annotated by COGs. This step leaves $M = 1108$ genomes.
- (2) Removal of COGs that are present in less than $\gamma_c = 0.4$ of the genomes, where gene frequencies are computed with sequence weights using $\delta = 0.9$. The relatively high value $\gamma_c = 0.4$ is meant to reduce the data to a size that is easily tractable computationally; here it leaves $L = 1474$ COGs. Conversely, the choice of $\delta = 0.9$ is meant to preserve a relatively high effective number of genomes, here $M' = 380$. The exact values of these parameters are however not crucial.

The data is represented by a $M \times L$ binary array x_{si} with $x_{si} = 1$ if genome s has at least one gene in COG i , and 0 otherwise. The average occurrence of genes is $q = \sum_{si} x_{si} / (ML) = 0.67$. This dataset is in no way meant to be optimal, and finer definitions of orthology are possible. Our point here is to show that sectors and sectons can be unraveled even from a relatively crude construction of bacterial phylogenetic profiles, leaving for future work the study of more elaborated datasets.

Genomic sectons are obtained exactly as protein sectons, except that the alphabet is now binary ($A = 1$), sequence weights are computed with $\delta = 0.9$, and \mathcal{D}_{ij} is not truncated ($\Delta = 0$). The content of first 100 sectons (obtained with $k_{\text{top}} = 120$) is reported in Fig. S20, with further details for the top 24 sectons provided in Fig. S23.

Genomic sectors can also be defined following the methods for defining protein sectors. Fig. S21 shows the counterpart of Fig. 1, using here $k_{\text{top}} = 6$. In absence of a counterpart for the experimentally determined three-dimensional structure of proteins, assessing the relevance of these sectors is not obvious. Using the partition of COGs into 3 broad functional classes [17], we nevertheless find that 4 of the 6 components support groups of COGs that are significantly enriched in some of these classes, as reported in Fig. S22. This suggests that genome sectors may be defined as well, which contain co-functional and therefore coevolving genes.

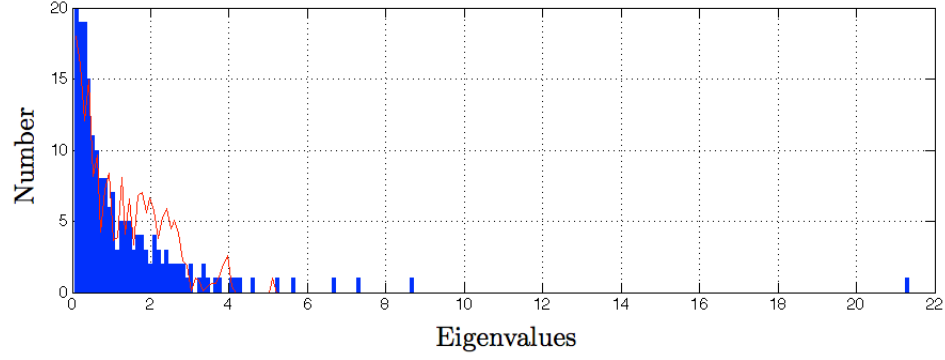


FIG. S 1: Spectrum of \mathcal{C}_{ij} for the trypsin family – In blue, histogram of the L eigenvalues of the matrix \mathcal{C}_{ij} (truncated to 20 along the y -axis). In red, average spectrum over 100 randomized alignments where the amino acids are drawn independently at each position i according to the frequencies f_i^a . This shows that between 4 and 8 eigenvalues may be considered as significant.

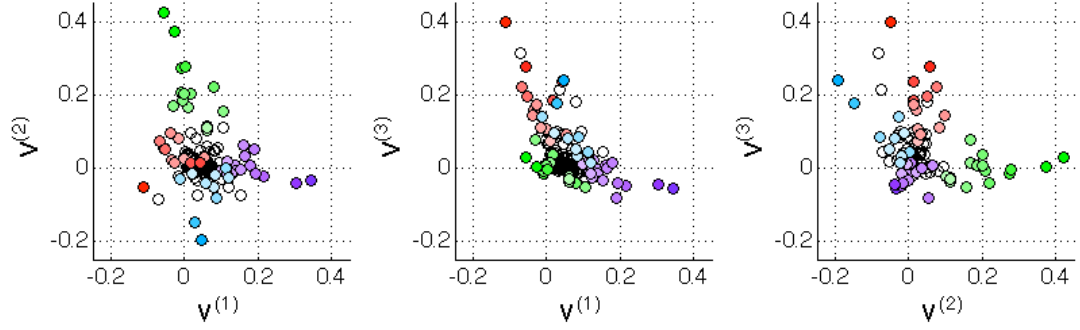


FIG. S 2: Independent components from rotation by ICA of the top $k_{\text{top}} = 3$ eigenvectors of \mathcal{C}_{ij} – This figure is the counterpart of Fig. 1, for $k_{\text{top}} = 3$ instead of $k_{\text{top}} = 4$, and with positions colored as in Fig. 1. The same green and red sectors are defined along $V^{(1)}$ and $V^{(2)}$, but the red and cyan sectors appear together along $V^{(3)}$.

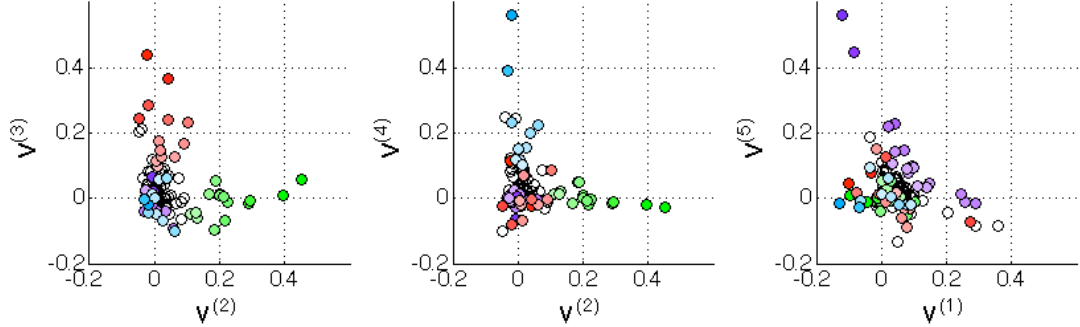


FIG. S 3: Independent components from rotation by ICA of the top $k_{\text{top}} = 5$ eigenvectors of \mathcal{C}_{ij} – This figure is the counterpart of Fig. 1, for $k_{\text{top}} = 5$ instead of $k_{\text{top}} = 4$, and with positions colored as in Fig. 1. The definitions of the green, red and cyan sectors along $V^{(2)}$, $V^{(3)}$, $V^{(4)}$ are consistent with their definition with $k_{\text{top}} = 4$, while the new component $V^{(5)}$ decomposes the purple sector in two subsets, whose interpretation is given in Fig. S4.

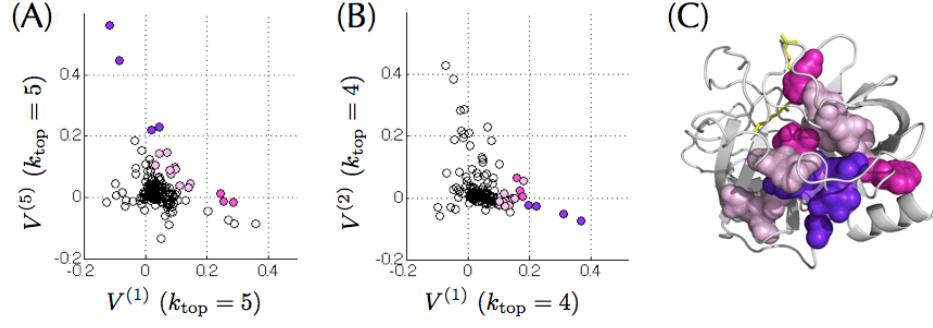


FIG. S 4: Interpretation of the decomposition of the purple sector when considering $k_{\text{top}} = 5$ – (A) Same as the last graph of Fig. S3, except that the positions from the purple sector are indicated with three different colors, depending on whether they satisfy $V_i^{(1)} > 0.2$ (pink), $V_i^{(5)} > 0.2$ (purple) or neither (light pink). (B) Using this coloring scheme but now for the projection along components obtained with $k_{\text{top}} = 4$ as in Fig. 1, showing that the partition of the purple sector corresponds to a partition between its core, defined as the positions with highest contribution to $V^{(1)}$, and the others. (C) Location of the positions on the three-dimensional structure, showing that the core is structurally connected with the other positions around it.

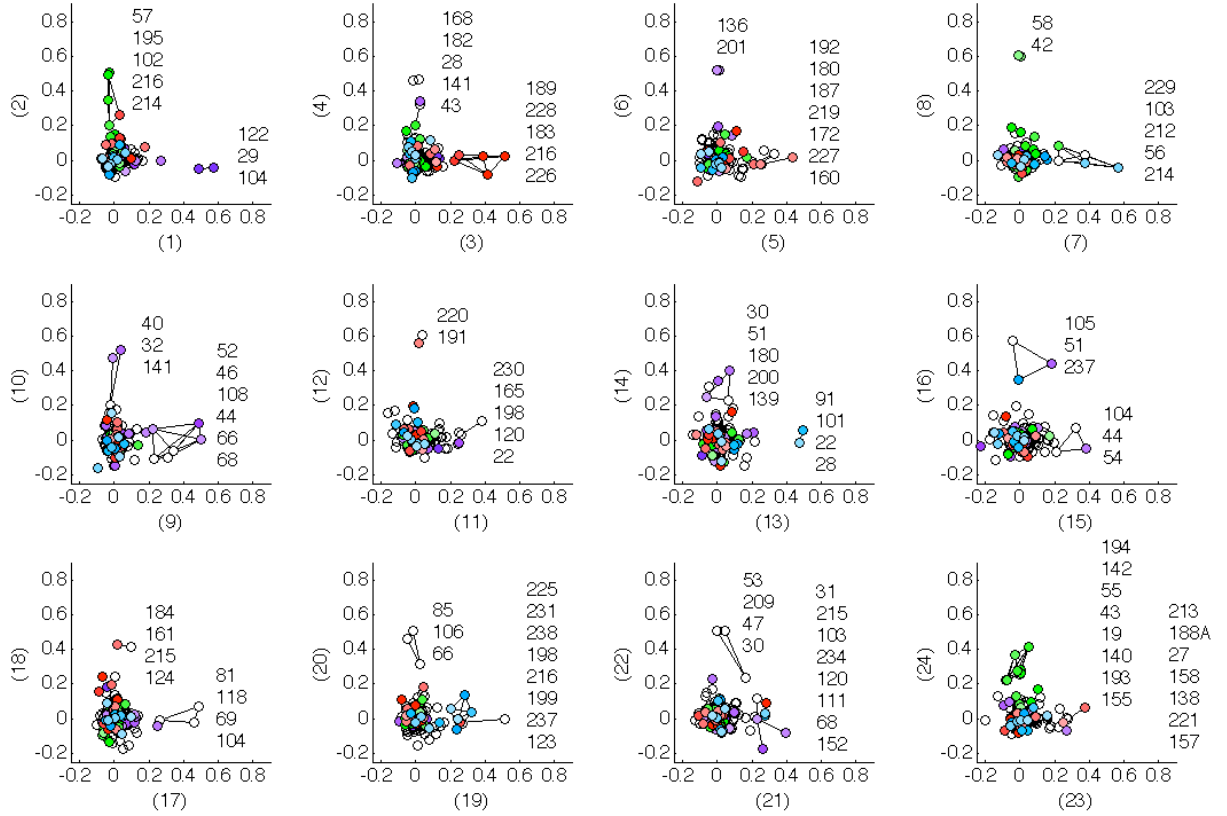


FIG. S 5: Independent components from rotation by ICA of the top $k_{\text{top}} = 24$ eigenvectors of C_{ij} – This figure is the counterpart of Fig. 1, for $k_{\text{top}} = 24$ instead of $k_{\text{top}} = 4$, and with positions colored as in Fig. 1. All sectors are now split up into smaller units that prefigure the sectors. Lines between the positions indicate structural contacts (for clarity, these contacts are represented only for positions with contribution > 0.2 along each component).

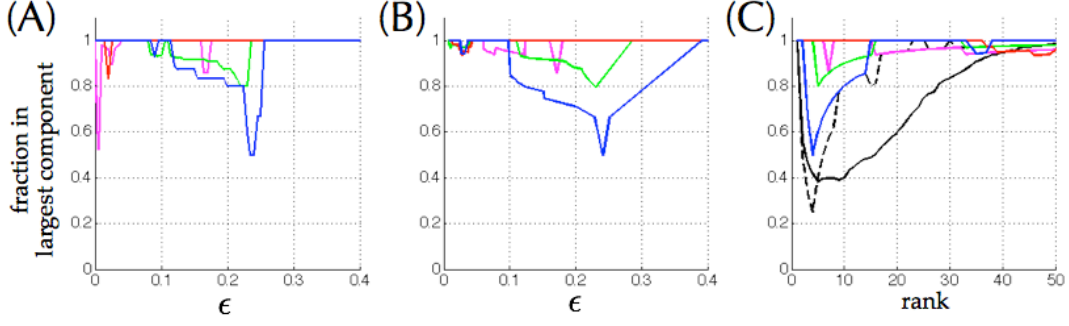


FIG. S 6: Connectivity of the sectors for varying values of ϵ – (A) Fraction of the positions in the largest structurally connected subset of a sector for varying values of ϵ and for sectors defined as in the main text by $\mathcal{S}'_k = \{i : V_i^{(k)} > \epsilon, V_i^{(\ell)} < \epsilon, \ell \neq k\}$ ($k = 1, \dots, k_{\text{top}} = 4$). The colors correspond to those of Fig. 1: magenta for $k = 1$, green for $k = 2$, red for $k = 3$ and blue for $k = 4$. (B) Same as (A) but not excluding intersections, which are significant for small ϵ , i.e., taking $\mathcal{S}_k = \{i : V_i^{(k)} > \epsilon\}$. (C) To assess the significance of (A) and (B), we consider also two other groups of ordered positions: randomly ranked positions and positions ranked by degree of conservation, measured by the relative entropy $D_i = \sum_{a=0}^{20} f_i^a \ln f_i^a / q^a$ with $f_i^0 = 1 - \sum_{a=1}^{20} f_i^a$ and $q^0 = 1 - \sum_{a=1}^{20} q^a$ for the frequencies of gaps. The results are presented here as a function of the size of the groups, where positions are added according to their rank (given by $V_i^{(k)}$ for the sectors). Compared to randomly ranked positions (full black line), sectors are clearly significantly more connected. They are also more connected than positions ranked by conservation (dashed black line), with the possible exception of the cyan sector (blue line).

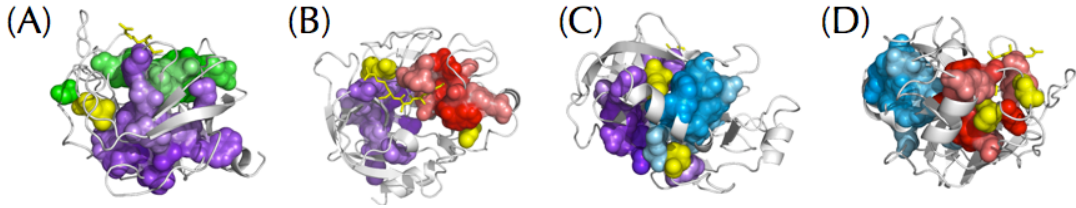


FIG. S 7: Intersections between \mathcal{S}_k – In the main text, sectors are defined by $\mathcal{S}'_k = \{i : V_i^{(k)} > \epsilon, V_i^{(\ell)} < \epsilon, \ell \neq k\}$ ($k = 1, \dots, k_{\text{top}} = 4, \epsilon = 0.1$). Here we represent on the three-dimensional structure the few positions that are excluded as ambiguous, i.e., belonging to intersections $\mathcal{S}_k \cap \mathcal{S}_\ell$ for $k \neq \ell$, with $\mathcal{S}_k = \{i : V_i^{(k)} > \epsilon\}$. There are 7 such positions, all of which are at the structural periphery of one of the two sectors \mathcal{S}'_k or \mathcal{S}'_ℓ , including 4 that are at the structural interface between the two. (A position from the green sector appears as disconnected in (A), but the position that could connect it to the rest of the green sector is actually not included in the alignment).

Purple sector	Green sector	Red sector	Cyan sector
29	57*	228*	237*
122	195*	189*	238
46*	197*	215*	234
120	55*	183*	91
28	19*	226*	231
51	102*	216*	103
40	43*	213*	229*
30	194*	227*	101
104*	193	184	123*
27	214*	191*	199
31	94	192*	
136*	142*	172*	
52*	42*	138	
201*	58*		
32	33*		
68*			
200			

FIG. S 8: Composition of the sectors defined in Fig. 1, ranked by the corresponding value of $V_i^{(k)}$ – A star indicates for the purple and cyan sectors that the position belongs to the blue sector defined in Ref. [5], and for the green and red sectors that it belongs respectively to the green and red sector of Ref. [5].

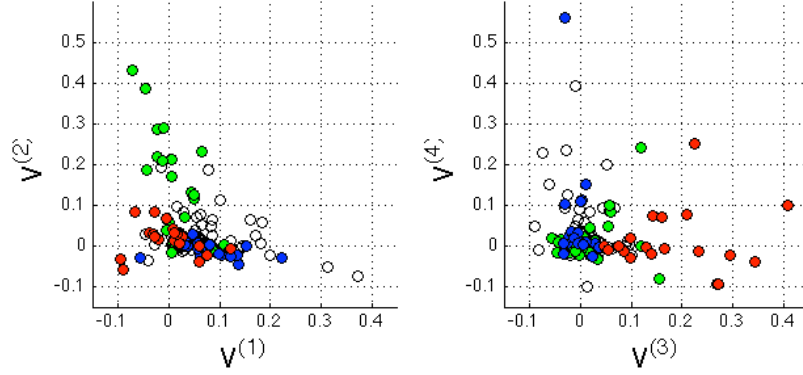


FIG. S 9: Comparison with the sectors defined in Ref. [5] – The graphs are identical to those of Fig. 1(A), except that the positions are here colored according to the definition of the 3 sectors of Ref. [5]. This shows as in Fig. S8 that essentially the same green and red sectors are identified, while the purple and cyan sectors have a small overlap with the previously defined blue sector.

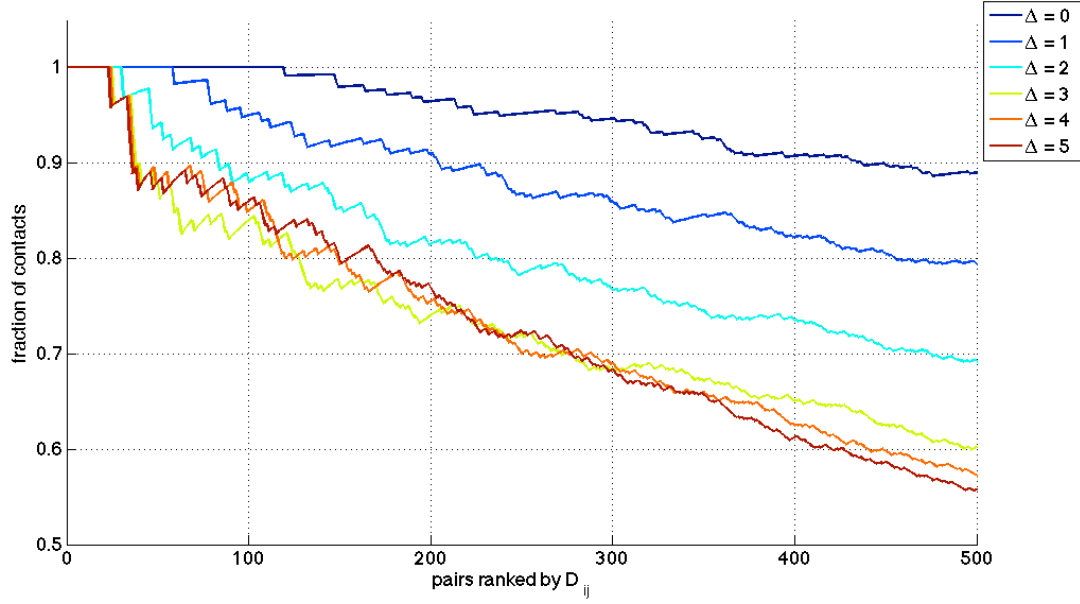


FIG. S 10: Fraction of pairs of positions predicted to be in contact from the top values of the matrix of direct information \mathcal{D}_{ij} – Pairs ij of positions are ordered by decreasing values of \mathcal{D}_{ij} , and the fraction of top n pairs to be in structural contact (distance $< 8 \text{ \AA}$) is considered as a function of n . Different curves correspond to different values of Δ , where pairs at distance $\leq \Delta$ along the sequence are ignored.

contact	section	dist. (\AA)	contact	section	dist. (\AA)	contact	section	dist. (\AA)	contact	section	dist. (\AA)
1	2	2	97	13	4	178	15	2.8	250	-	6.7
2	3	2	108	29	4.4	179	-	3.8	255	27	6.6
3	5	2.2	109	26	3	182	30	4.1	256	-	4.2
4	4	2.7	120	-	21	185	23	3.4	262	19	3.1
5	6	2	121	-	2	191	-	3.6	264	116	2.6
6	9	2.8	123	-	3.5	193	20	3.7	269	-	9.9
7	8	6.7	127	24	2.7	196	-	13	271	-	3.3
12	7	4.9	128	16	3.5	197	30	2.6	274	41	2
18	7	3.8	133	25	3.1	207	20	4.7	279	20	4.2
28	12	3.9	136	16	3.8	211	-	4	280	51	4.7
39	10	3.8	138	51/61	3.9	216	-	4.5	281	-	3.6
40	14	2.7	139	13	3.8	218	28	3.9	282	-	12
41	11	3.3	140	17	3.6	223	-	12	286	-	7.7
43	4	2.7	148	23	26	228	38	3.8	288	83	28
49	10	3.6	149	-	21	231	-	2.8	289	-	5.4
52	18	3.6	164	-	22	233	34	3.5	291	26	3
54	15	3.8	165	32	3.5	234	-	3.6	292	58	3.8
78	21	2.7	166	21	3.3	236	36	3.6	294	13	3.9
86	19	4.3	176	-	11	239	33	2.9	297	-	3.3
96	22	3.3	177	31	5	243	76	3.8	298	-	3.8

FIG. S 11: Top 80 non-trivial contacts and their associated section – The pairs are ordered by decreasing values of \mathcal{D}_{ij} , with their rank indicated in the first column. Only pairs of positions at distance $> \Delta = 5$ along the sequence are considered, and many pairs are therefore not included (red curve in Fig. S10). The second column indicates the rank of the section where the pair is found and the third the distance between the positions in the three-dimensional structure. Pairs are considered in contact when this distance is $< 8 \text{ \AA}$, and false positive are indicated in red. Note that most of these false positive are actually not in a section: if considering the 56 pairs that both in one of the top 80 non-trivial pairs and in one of the top 120 sections (all those shown here), only 2 are false positive (96% positive rate) while out of the top 56 non-trivial pairs 7 are false positive (87% positive rate); this suggests that integrating sections into the prediction of contacts may lead to an increased positive rate.

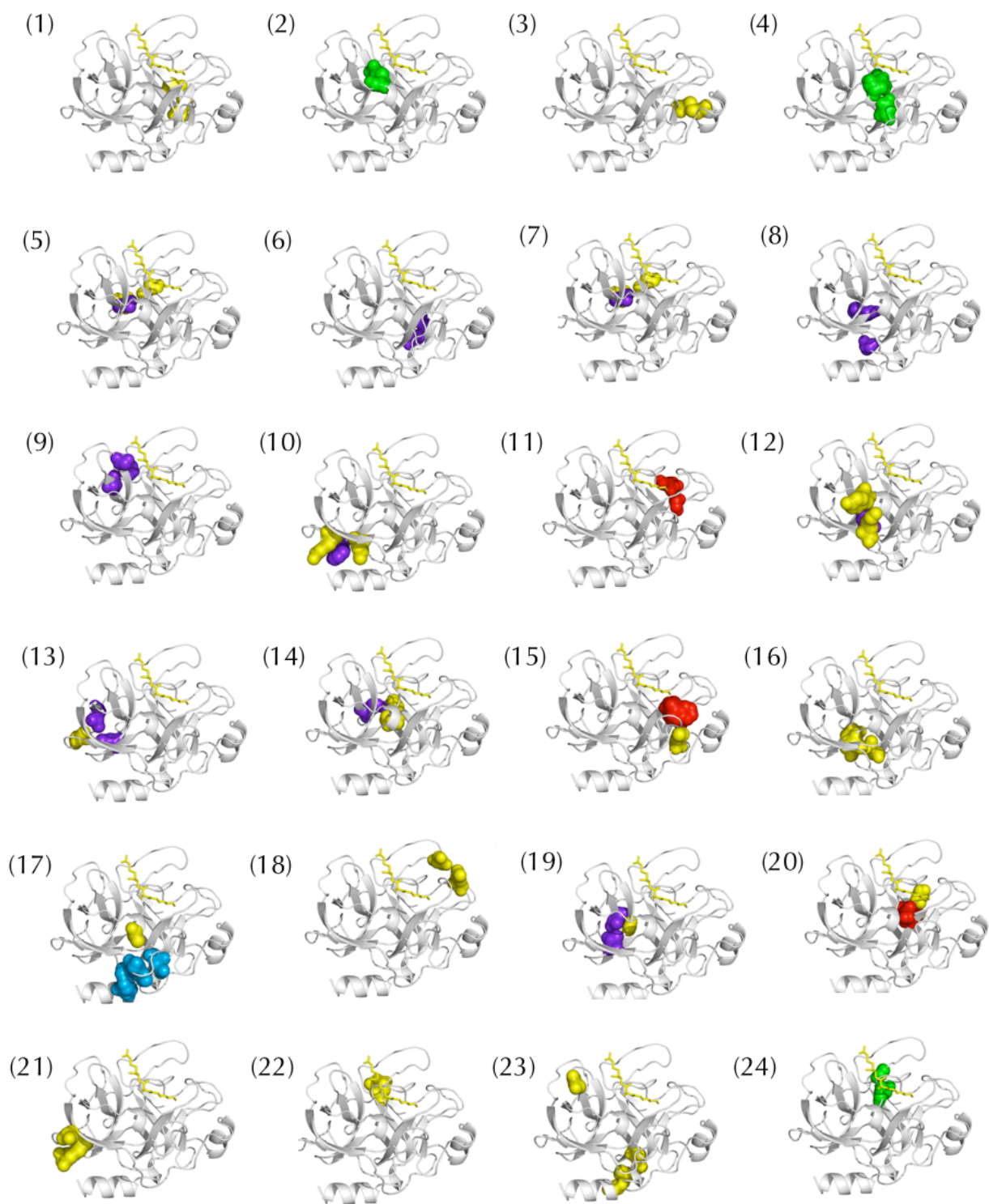


FIG. S 12: Structural representation of the top 24 sectors – Representation on the three-dimensional structure of rat trypsin of the top 24 sectors defined in Fig. 2. The colors refer to the sectors, with yellow for non-sector positions. Only section 23 contains a position that is disconnected from the others when taking a distance $< 8\text{\AA}$ as criterion for physical connectivity.

Sectons	Positions	Sectons	Positions	Sectons	Positions	Sectons	Positions
*1	159 135	*31	165 230 176	*61	221 228 189	*91	204
*2	58 42	*32	158 188A 138	*62	123 235	*92	38
*3	168 182	*33	109 84 83	*63	33 41 66	93	129 203
*4	57 195 102	*34	67 82 34	*64	214 102	*94	23
*5	191 220	*35	85 106 66	*65	74 153	*95	93 99
*6	201 136	*36	111 49 50 83	*66	108 83 66	96	28 119 69 39
*7	157 26 27	*37	78 72	*67	197 43 55	*97	73
*8	29 122	*38	118 81	68	216 227 210	*98	88 106
*9	32 40	*39	92 237	*69	178 169 172	*99	19
*10	51 105 107	*40	121 45 200	*70	148	100	90 99
*11	226 189	41	232 128 22	*71	173	*101	47 238 53
*12	104 59 90	42	163 225 119	*72	110 83	*102	151 143
*13	112 46 68	*43	192 143 219 221	*73	231 238	*103	140
*14	30 139 198	*44	144 152	*74	170	*104	24 117
*15	215 180 227	*45	25 117 155	75	134 125	105	116 233
*16	209 53 45 121	*46	77 70	*76	184A 161	*106	236
*17	91 101 237 56	*47	229 212 103	*77	95	*107	130
*18	221A 146	*48	69 141	*78	64	*108	198
*19	44 52 31	*49	56 94	*79	154	*109	147
*20	190 213 138	*50	162 210 133	*80	115 81	*110	61
*21	107 86 50	*51	228 183 199 160 189	81	54 166	111	166 172 49
*22	156 21	52	34 155 141 22 28	*82	174	*112	176 169
23	124 37 235 210	53	133 50	83	113 101	*113	48
*24	142 194	*54	132 164	*84	76	*114	193 43
*25	184 161	*55	98	85	60 94	*115	202
*26	179 100 233	56	89 99 153	*86	96	116	177 100 171
*27	181 199	*57	222 187	*87	167	*117	63 106
*28	114 120	58	31 39 68	*88	196	*118	55 212 197
*29	137 27	*59	224 217	89	170 151	*119	188
*30	234 103 101 229	60	160 43 219 172 228 138	*90	211	120	87 107 93

FIG. S 13: Composition of sections – Sections are defined here with $k_{\text{top}} = 120$ and $\epsilon = 0.2$. A star indicates that the section is structurally connected. Up to rank 40, all but one section are thus connected. Note that at large rank, some sections consist of single positions and are therefore trivially connected; see also Fig. S14.

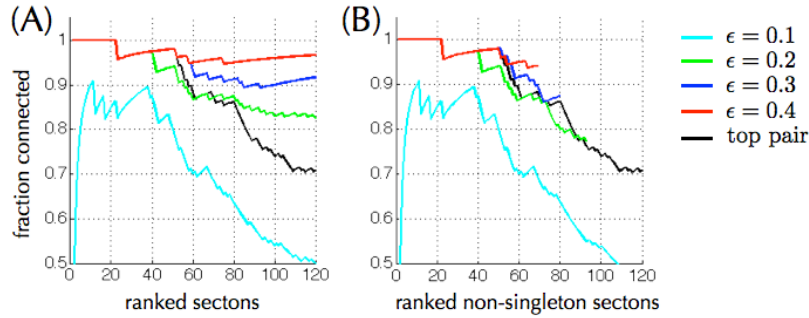


FIG. S 14: Connectivity of sections – Sections are here defined for $k_{\text{top}} = 120$ and different values of ϵ . A section is considered as structurally connected if all the positions that it contains are in contact in the three-dimensional structure, either directly or indirectly. **(A)** Fraction of the top sections to be structurally connected as function of the rank up to which sections are considered. For comparison, the black curve corresponds to the top 2 positions contributing to $U^{(k)}$. **(B)** Same as (A) but excluding the sections that contain a single position and are therefore trivially connected. These figures show that by considering a larger, more stringent threshold ϵ than in Fig. S13 (where $\epsilon = 0.2$) more (but smaller) sections are found to be connected.

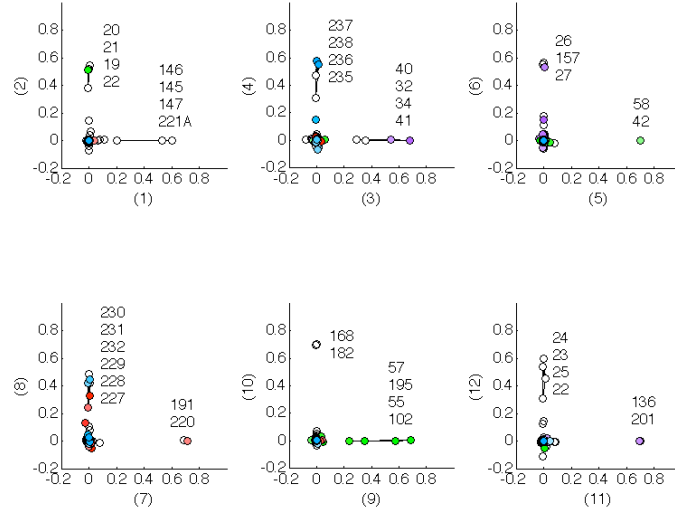


FIG. S 15: Sectors when not truncating \mathcal{D}_{ij} to $\tilde{\mathcal{D}}_{ij}$ – Out of the 24 sectors displayed here, 5 consist exclusively of consecutive positions and may be attributed to consecutive gaps.

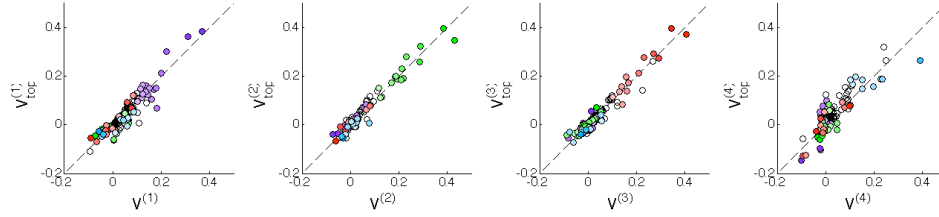


FIG. S 16: Sectors from the top 20 modes of \bar{C} – We compare here the top $k_{\text{top}} = 4$ independent components derived from C and \bar{C}^+ to show that they are highly correlated and therefore define the same sectors.

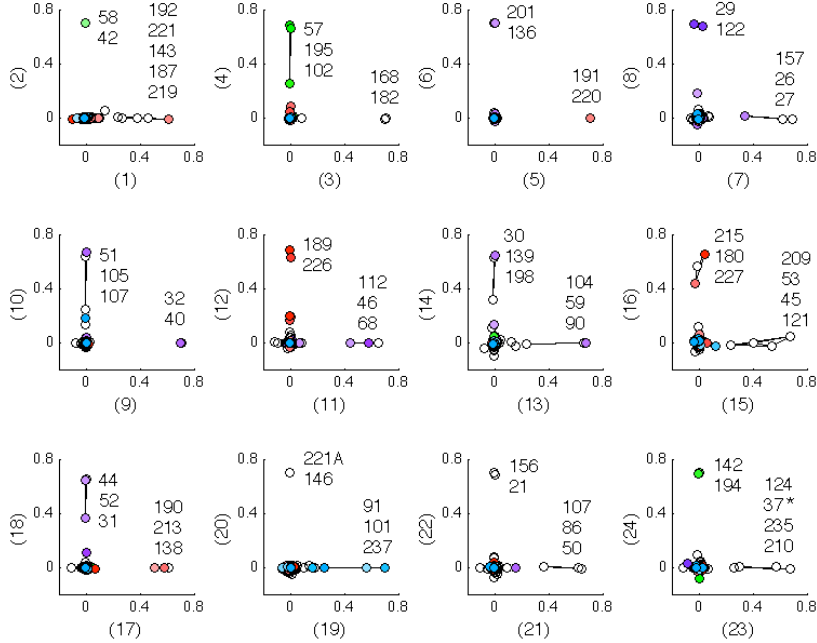


FIG. S 17: Sectors from the bottom modes of \bar{C} – We show here the top 24 sectors derived from \bar{C}^- . A comparison with Fig. 2 shows that essentially the same sectors are defined by this matrix and by \bar{C} .

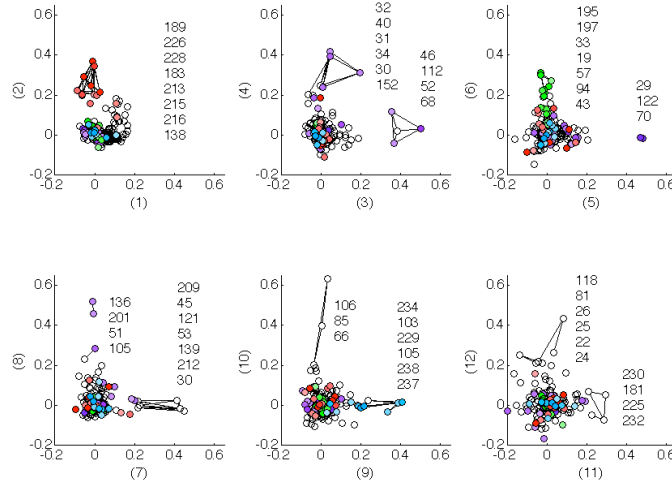


FIG. S 18: Sector analysis from the unweighted correlation matrix C_{ij}^0 – Uniform positional weights are used ($W_i^a = 1$ for all i and a) and the top $k_{\text{top}} = 12$ components are considered. Lines between the positions indicate structural contacts (for clarity, these contacts are represented only for the top positions along each component). The group of coevolving positions defined by the components are intermediate between sectors and sections. As far as the top 6 components are concerned, components $k = 2$ and $k = 6$ thus correspond to the cores of the red and green sectors while components 3, 4 and 5 correspond to the sections 13, 9 and 8 defined in Fig. 2, but with one or more additional positions (component 1 is here not localized).

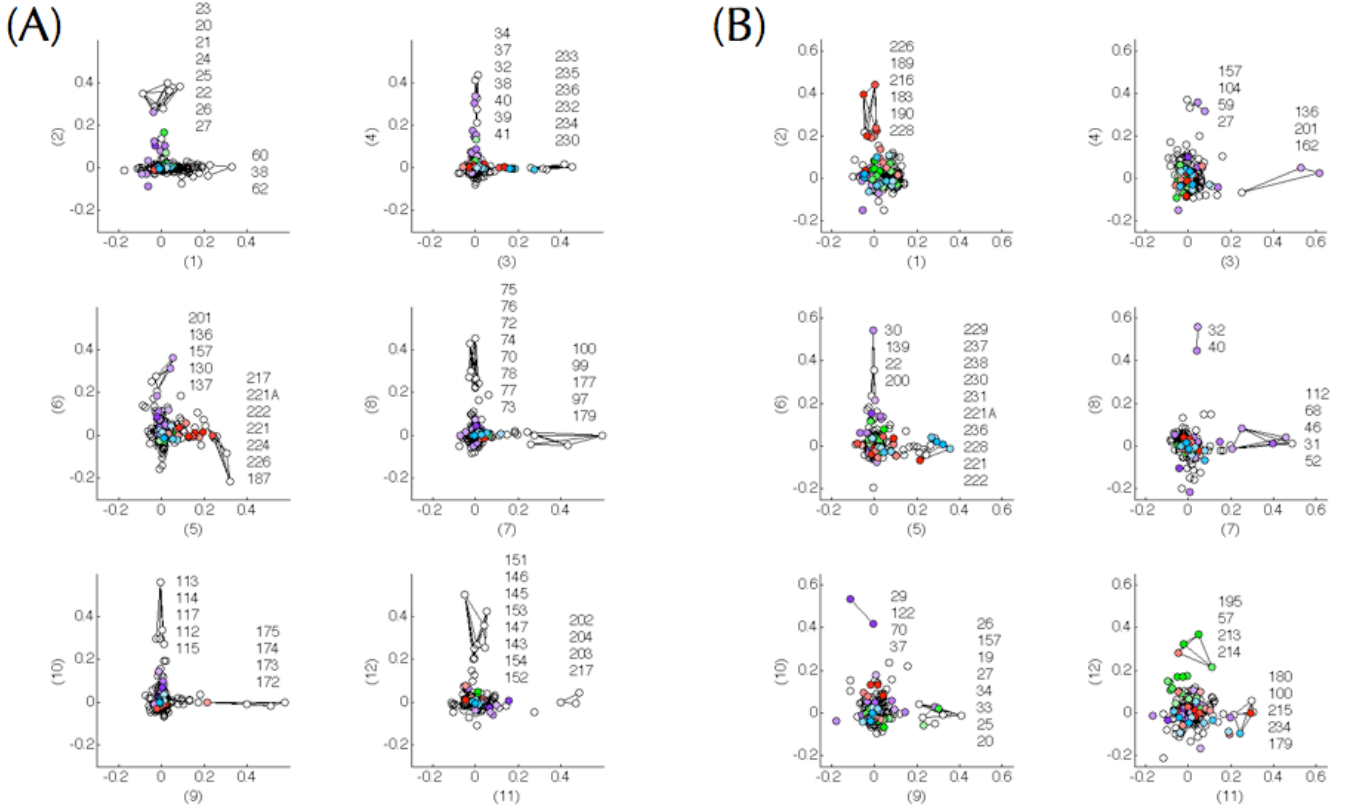


FIG. S 19: Sector analysis from the matrix of mutual information \mathcal{M}_{ij} – **(A)** Similar to Fig. S18 but using \mathcal{M}_{ij} instead of C_{ij}^0 . **(B)** After truncating \mathcal{M}_{ij} from a band in its diagonal, i.e. using $\tilde{\mathcal{M}}_{ij}$ with $\tilde{\mathcal{M}}_{ij} = \mathcal{M}_{ij}$ when $|i - j| > \Delta = 5$ and 0 otherwise. The truncation reveals more structurally connected groups of positions that are not consecutive.

section	COGs
1	
2	1006P 1863P 1320P 651CP
3	1984E 2049E 1540R
4	559E 410E 411E 4177E 683E
5	4603R 3845R 1079R 1744R
6	1135P 2011P 1464P
7	1010H 1797H 2875H 2082H 2242H 2243H
8	1319C 1975O 2068R 2080C 1529C
9	1176E 1177E 687E
10	286V 610V 732V
11	1005C 1007C 1008C 649C 713C 838C 839C 377C
12	1346M 1380R
13	1788I 2057I
14	1271C 1294C
15	163H 43H
16	1732M 1125E
17	1638G 1593G
18	1203R 1518L
19	1653G 3839G 395G 1175G
20	2894D 850D 851D
21	2332O 2386O 1138O
22	1129G 1172G 1879G
23	3451U 3505U
24	1640G 58G 296G 297G 1523G 448G
25	1116P 600P 715P
26	1228Q 2986E 2987E
27	1270H 2038H 2087H 1492H 368H
28	1088M 1091M 1209M 1898M
29	3288C 1282C
30	22C 1071C 508C
31	396O 719O
32	1677NU 1684NU 1766NU [...]
33	1622C 109O 1845C 843C 1612O
34	4962U 4965U
35	1918P 370P
36	241E 279G 2870M 859M
37	117H 1985H
38	2804NU 1989NOU 1450NU 1459NU
39	1120PH 609P 614P
40	2896H 303H 746H 521H 314H 315H
41	1178P 1840P
42	1352NT 643NT 2201NT 835NT 840NT
43	2025C 2086C
44	1034C 1894C 1905C
45	1003E 404E 509E
46	1127Q 1463Q 767Q
47	422H 2022H 2104H
48	379H 29H 157H
49	1366T 2172T 2208TK
50	2224C 2225C

section	COGs
51	5405O 1220O
52	1173EP 601EP 747E 444EP
53	1682GM 1134GM
54	226P 1117P 573P 581P
55	280C 282C
56	1108P 1121P 803P
57	1663M 1043M 1212M [...]
58	1126E 765E 834ET
59	1392P 306P
60	2884D 2177D
61	1290C 2010C 723C
62	588G 696G
63	2009C 1485R 2142C
64	175EH 2895P 7H 155P
65	2801L 2963L
66	1122P 619P
67	45C 74C
68	1828F 47F
69	1354S 1386K
70	106E 107E 118E 40E 131E 139E 141E
71	1183I 688I
72	168P 569P
73	1013C 674C
74	1080G 1762GT 1925G
75	1180O 1328F
76	132H 502H 156H 161H
77	751J 752J
78	3275T 3279KT
79	1704S 1512R
80	1706N 1261NO 2063N
81	245I 1154HI 1211I 1947I 743I 761IM 821I
82	1020Q 2091H 736I
83	1118P 555O
84	4799I 511I 777I 825I
85	3383R 437C 1526C
86	602O 603R 720H
87	263E 14E
88	554C 578C
89	113H 181H 1H 7H 373H
90	413H 414H 853H
91	2805NU 1989NOU 4972NU
92	1837R 1847R
93	547E 133E 134E 135E 147EH 512EH 159E
94	3206M 1596M
95	1191K 1317NU 1345N 1516NUO 1551T 3144N
96	1077D 1792M 1426S
97	246G 2723G 2017G 800G
98	717F 2131F
99	4149P 725P
100	419L 420L

FIG. 20: Content of the top 100 genomic sections in terms of COGs – The full content of section 32 is {1677NU 1684NU 1766NU 1157NU 1815N 1843N 1868N 1256N 1291N 1298NU 1338NU 1344N 1987NU 1377NU 1536N 1558N} and the full content of section 57 is {1663M 1043M 1212M 2877M 763M 774M 1519M 794M 1560M}.

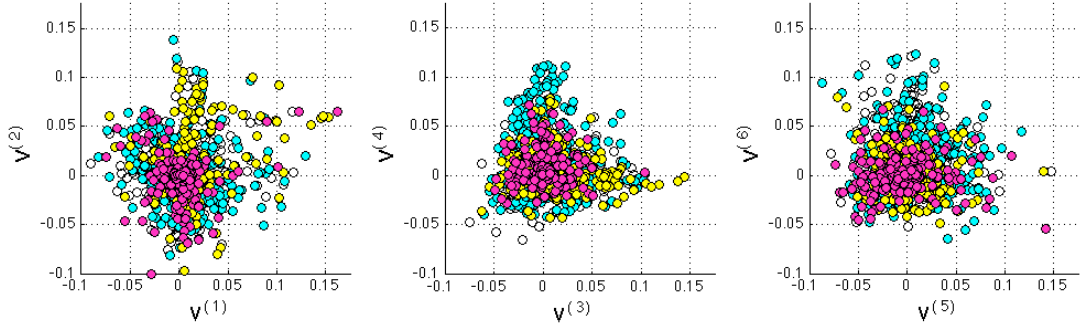


FIG. S 21: Genomic sectors ($k_{\max} = 6$) – COGs are colored as in Fig. 3 according to the functional category to which they belong: cyan for metabolism, yellow for cellular processes, magenta for information processing, and white for poorly characterized genes or genes that belong to multiple categories [17]. The apparent enrichment of yellow COGs along $V^{(2)}$ or cyan COGs along $V^{(4)}$ is quantitatively estimated in Fig S22.

Sector	Total number of genes	Info. process.	Cellular process.	Metabolism	p -value for the composition
1	22	1	4	13	0.14
2	56	5	25	17	$5 \cdot 10^{-5}$
3	74	2	35	23	$5 \cdot 10^{-9}$
4	114	8	10	85	$3 \cdot 10^{-9}$
5	50	12	7	24	0.37
6	65	2	4	32	$3 \cdot 10^{-4}$

FIG. S 22: Association between genomic sectors and functional categories – For each sector, defined as the COGs with contribution > 0.05 along one of the 6 components $V^{(k)}$ shown in Fig.S 22 and not along any other, we assessed the significance of their content in information processing (magenta), cellular processing (yellow), metabolic (cyan) COGs, which represent respectively around 1/2, 1/4 and 1/4 of the COGs. The p -value is computed from a χ^2 -square test with 2 degrees of freedom. 4 of the 6 sectors can be considered to be significantly enriched in COGs of some of these functional categories.

Secton	COG	Annotation
2	1006P	Multisubunit Na ⁺ /H ⁺ antiporter, MnhC subunit
	1863P	Multisubunit Na ⁺ /H ⁺ antiporter, MnhE subunit
	1320P	Multisubunit Na ⁺ /H ⁺ antiporter, MnhG subunit
	651CP	Formate hydrogenlyase subunit 3/Multisubunit Na ⁺ /H ⁺ antiporter, MnhD subunit
3	1984E	Allophanate hydrolase subunit 2
	2049E	Allophanate hydrolase subunit 1
	1540R	Uncharacterized proteins, homologs of lactam utilization protein B
4	559E	Branched-chain amino acid ABC-type transport system, permease components
	410E	ABC-type branched-chain amino acid transport systems, ATPase component
	411E	ABC-type branched-chain amino acid transport systems, ATPase component
	4177E	ABC-type branched-chain amino acid transport system, permease component
	683E	ABC-type branched-chain amino acid transport systems, periplasmic component
5	4603R	ABC-type uncharacterized transport system, permease component
	3845R	ABC-type uncharacterized transport systems, ATPase components
	1079R	Uncharacterized ABC-type transport system, permease component
	1744R	Uncharacterized ABC-type transport system, periplasmic component/surface lipoprotein
6	1135P	ABC-type metal ion transport system, ATPase component
	2011P	ABC-type metal ion transport system, permease component
	1464P	ABC-type metal ion transport system, periplasmic component/surface antigen

FIG. S 23: Annotations for the top genomic sections – Note that there is no section along the first component, as seen in Fig. 3. (Sections 7 to 24 are presented on the next page). This shows that the definition of sections is consistent with our current knowledge of gene functions.

7	1010H	Precorrin-3B methylase
	1797H	Cobyrrinic acid a,c-diamide synthase
	2875H	Precorrin-4 methylase
	2082H	Precorrin isomerase
	2242H	Precorrin-6B methylase 2
	2243H	Precorrin-2 methylase
8	1319C	Aerobic-type carbon monoxide dehydrogenase, middle subunit CoxM/CutM homologs
	1975O	Xanthine and CO dehydrogenases maturation factor, XdhC/CoxF family
	2068R	Uncharacterized MobA-related protein
	2080C	Aerobic-type carbon monoxide dehydrogenase, small subunit CoxS/CutS homologs
	1529C	Aerobic-type carbon monoxide dehydrogenase, large subunit CoxL/CutL homologs
9	1176E	ABC-type spermidine/putrescine transport system, permease component I
	1177E	ABC-type spermidine/putrescine transport system, permease component II
	687E	Spermidine/putrescine-binding periplasmic protein
10	286V	Type I restriction-modification system methyltransferase subunit
	610V	Type I site-specific restriction-modification system, R (restriction) subunit and related helicases
	732V	Restriction endonuclease S subunits
11	1005C	NADH:ubiquinone oxidoreductase subunit 1 (chain H)
	1007C	NADH:ubiquinone oxidoreductase subunit 2 (chain N)
	1008C	NADH:ubiquinone oxidoreductase subunit 4 (chain M)
	649C	NADH:ubiquinone oxidoreductase 49 kD subunit 7
	713C	NADH:ubiquinone oxidoreductase subunit 11 or 4L (chain K)
	838C	NADH:ubiquinone oxidoreductase subunit 3 (chain A)
	839C	NADH:ubiquinone oxidoreductase subunit 6 (chain J)
	377C	NADH:ubiquinone oxidoreductase 20 kD subunit and related Fe-S oxidoreductases
12	1346M	Putative effector of murein hydrolase
	1380R	Putative effector of murein hydrolase LrgA
13	1788I	Acyl CoA:acetate/3-ketoacid CoA transferase, alpha subunit
	2057I	Acyl CoA:acetate/3-ketoacid CoA transferase, beta subunit
14	1271C	Cytochrome bd-type quinol oxidase, subunit 1
	1294C	Cytochrome bd-type quinol oxidase, subunit 2
15	163H	3-polyprenyl-4-hydroxybenzoate decarboxylase
	43H	3-polyprenyl-4-hydroxybenzoate decarboxylase and related decarboxylases
16	1732M	Periplasmic glycine betaine/choline-binding (lipo)protein of an ABC-type transport system
	1125E	ABC-type proline/glycine betaine transport systems, ATPase components
17	1638G	TRAP-type C4-dicarboxylate transport system, periplasmic component
	1593G	TRAP-type C4-dicarboxylate transport system, large permease component
18	1203R	Predicted helicases
	1518L	Uncharacterized protein predicted to be involved in DNA repair
19	1653G	ABC-type sugar transport system, periplasmic component
	3839G	ABC-type sugar transport systems, ATPase components
	395G	ABC-type sugar transport system, permease component
	1175G	ABC-type sugar transport systems, permease components
20	2894D	Septum formation inhibitor-activating ATPase
	850D	Septum formation inhibitor
	851D	Septum formation topological specificity factor
21	2332O	Cytochrome c-type biogenesis protein CcmE
	2386O	ABC-type transport system involved in cytochrome c biogenesis, permease component
	1138O	Cytochrome c biogenesis factor
22	1129G	ABC-type sugar transport system, ATPase component
	1172G	Ribose/xylose/arabinose/galactoside ABC-type transport systems, permease components
	1879G	ABC-type sugar transport system, periplasmic component
23	3451U	Type IV secretory pathway, VirB4 components
	3505U	Type IV secretory pathway, VirD4 components
24	1640G	4-alpha-glucanotransferase
	58G	Glucan phosphorylase
	296G	1,4-alpha-glucan branching enzyme